FoDL priority program

DFG

MAX PLANCK INSTITUTE
FOR MATHEMATICS IN THE SCIENCES

# COMBINATORIAL AND ALGEBRAIC APPROACHES TO DEEP LEARNING

**Marie-Charlotte Brandenburg and Angélica Torres**

**Annual Meeting: Theoretical Foundations of Deep Learning**
**6 November 2023**

**Project:** Combinatorial and Implicit Approaches to Deep Learning

**Members:** Marie-Charlotte Brandenburg, Guido Montúfar, Johannes Müller, Angélica Torres, Hanna Tseran, Bernd Sturmfels

# OVERVIEW

## Part I: Combinatorial Approaches to Deep Learning

- Motivation: Parameter space of Linear Classifiers
- Parameter Space of ReLU Classifiers

## Part II: Algebraic Approaches to Deep Learning

- Motivation: Dynamics of gradient descent
- Polynomial invariances of a NN when optimizing its parameters using gradient descent

# Part I: Combinatorial Approaches to Deep Learning

Georg Loho
FU Berlin | University of Twente

Guido Montúfar
UC LA | MPI MiS

# LINEAR CLASSIFIERS

## Setup

Given data points $D = \{p_1, \ldots, p_n\} \in \mathbb{R}^d$,
a linear classifier is a linear function $f : \mathbb{R}^d \to \mathbb{R}$

- $f$ defines a hyperplane $\{x \in \mathbb{R}^d \mid f(x) = 0\}$ in input space separating $\{p_i \mid f(p_i) > 0\}$ from $\{p_i \mid f(p_i) < 0\}$
- $f$ can be parametrized as $f(x) = \langle a, x \rangle + b$ for some fixed $a \in \mathbb{R}^d, b \in \mathbb{R}$.
- parameter space of linear classifiers is $\{(a, b) \mid a \in \mathbb{R}^d, b \in \mathbb{R}\} \cong \mathbb{R}^{d+1}$.

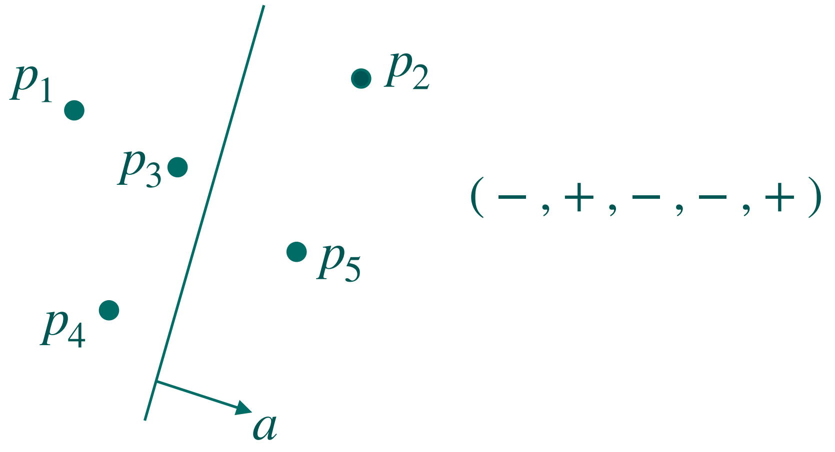
Classification by $f$:
(sgn($f(p_1)$), …, sgn($f(p_n)$)) $\in \{-,0,+\}^n$

## Goal

Subdivide parameter space into cells, in which classifiers have the same classification

## Theorem

These cells are chambers in the hyperplane arrangement $\bigcup_{p \in D} (p,1)^\perp \subseteq \mathbb{R}^{d+1}$



# LINEAR CLASSIFIERS

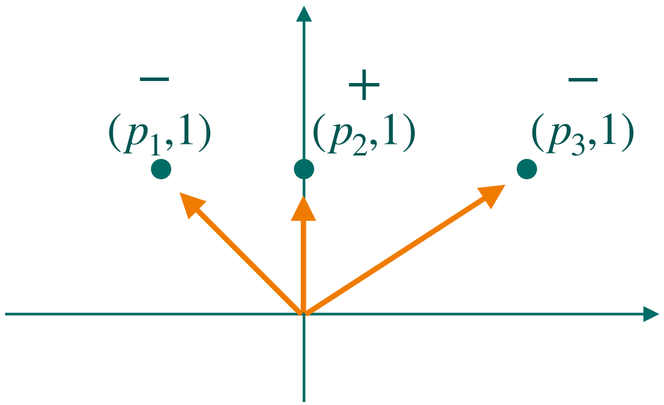
Fix a labelling $D = D^+ \sqcup D^-$

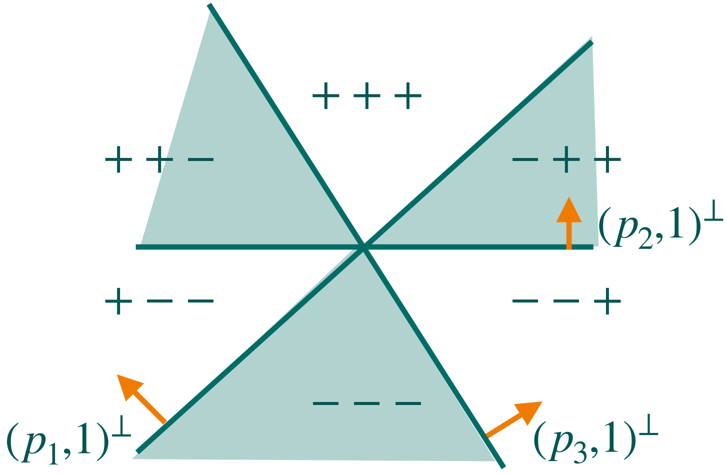
$f$ makes a mistake at $p \in D^+$ if $f(p) < 0$
$f$ makes a mistake at $p \in D^-$ if $f(p) > 0$

$0/1-$loss counts number of mistakes of $f$

**Proposition:**
All local minima are global minima.
More precisely, for any chamber $D$ there exist a chamber $C$ with minimum number of mistakes and a sequence $D = D_0, D_1, \dots, D_k, D_{k+1} = C$ such that $D_i, D_{i+1}$ are connected through codimension $1$ and the number of mistakes is strictly decreasing.

CAN WE GENERALIZE THIS TO LARGER CLASSES OF CLASSIFIERS?



# RELU NNS AND TROPICAL GEOMETRY

$D \subseteq \mathbb{R}^d$ data points, classified by a ReLU NN

**Theorem [Arora-Basu-Mianjy-Mukherjee '18]:**

Every ReLU NN represents a piecewise linear function, and every piecewise linear function $f : \mathbb{R}^d \to \mathbb{R}$ can be represented by a ReLU NN with at most $\lceil \log_2(d+1) \rceil + 1$ depth.

**Theorem [Zhang-Naitzat-Lim '18]:**

Every ReLU NN represents a tropical rational function, and every tropical rational function $f = g \oslash h$ can be represented by a ReLU NN with at most $\max(\lceil \log_2(n) \rceil, \lceil \log_2(m) \rceil) + 2$ depth, where $n, m$ are the number of monomials of $g, h$ respectively.

**— Tropical Intermezzo —**

$$a \oplus b = \max(a, b), \ \ a \odot b = a + b, \ \ a \oslash b = a - b, \ \ x^{\odot a} = a \cdot x$$

classical rational function

$$\tilde{f}(x) = \left( \textstyle\sum_{i=1}^{n} a_i x_1^{s_{i1}} \cdots x_d^{s_{id}} \right) \Big/ \left( \textstyle\sum_{j=1}^{m} b_j x_1^{t_{j1}} \cdots x_d^{t_{jd}} \right)$$

tropical rational function

$$f = \left( \bigoplus_{i=1}^{n} a_i \odot x_1^{\odot s_{i1}} \odot \ldots \odot x_d^{\odot s_{id}} \right) \oslash \left( \bigoplus_{j=1}^{m} b_j \odot x_1^{\odot t_{j1}} \odot \ldots \odot x_d^{\odot t_{jd}} \right)$$

$$= \max_{i=1,\ldots,n} \left( a_i + s_{i1}x_1 + \ldots + s_{id}x_d \right) - \max_{j=1,\ldots,m} \left( b_j + t_{j1}x_1 + \ldots + t_{jd}x_d \right)$$

$$= \max_{i=1,\ldots,n} \left( a_i + \langle s_i, x \rangle \right) - \max_{j=1,\ldots,m} \left( b_j + \langle t_j, x \rangle \right), \quad a_i, b_j \in \mathbb{R}, \ s_i, t_j \in \mathbb{R}^d$$

$=$ difference of two convex piecewise linear functions

$(n, m) = (1,1)$ recovers linear classifiers

# DECISION BOUNDARIES OF RELU NNS

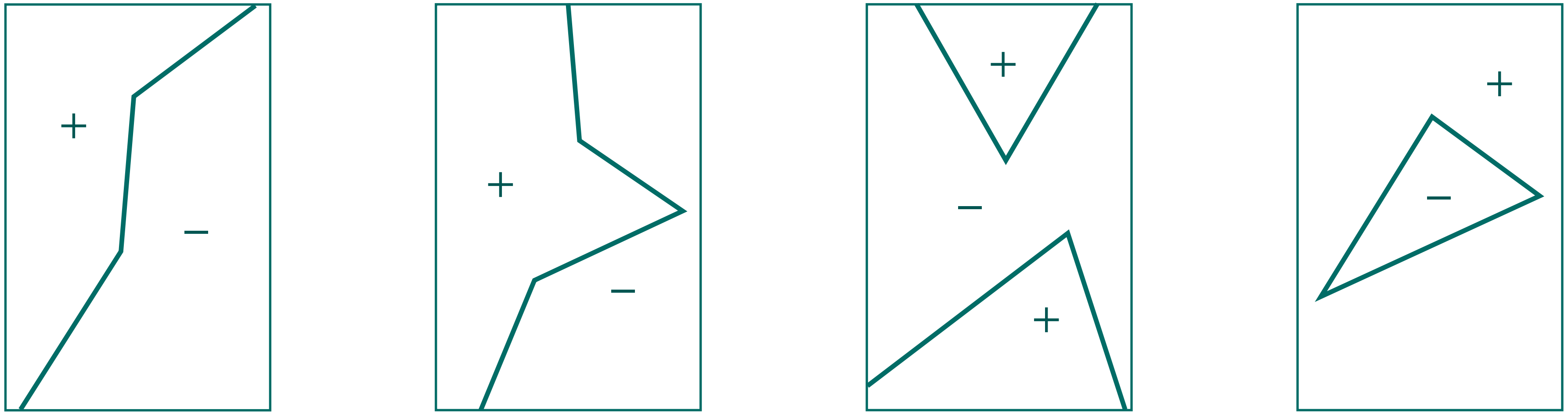
ReLU NNs: $f(x) = \max_{i=1,\dots,n} \left(a_i + \langle s_i, x\rangle\right) - \max_{j=1,\dots,m} \left(b_j + \langle t_j, x\rangle\right)$

Decision boundary $\{x \in \mathbb{R}^d \mid f(x) = 0\}$

Linear classifiers: hyperplanes

ReLU: Polyhedral complexes with at most $n \cdot m$ linear pieces

input dimension
$d = 2$




$(n, m) = (2,2)$ $(n, m) = (2,2)$ $(n, m) = (2,2)$ $(n, m) = (3,1)$

# SUBDIVISION OF PARAMETER SPACE

ReLU NNs: $f(x) = \max_{i=1,\ldots,n} \left(a_i + \langle s_i, x\rangle\right) - \max_{j=1,\ldots,m} \left(b_j + \langle t_j, x\rangle\right)$

Parameter space of tropical rational functions with $(n, m)$ terms:

$$\{\theta = (a_1, s_s, \ldots, a_n, s_n, b_1, t_1, \ldots, b_m, t_m) \mid a_i, b_i \in \mathbb{R}, s_i, t_j \in \mathbb{R}^d\} \cong \mathbb{R}^{(m+n)(d+1)}$$

Subdivide parameter space into cells, where classifiers have the same classification:

for fixed labelling $D = D^+ \sqcup D^-$, consider $\theta = (a_1, s_s, \ldots, a_n, s_n, b_1, t_1, \ldots, b_m, t_m)$ such that

$$\max_{i=1,\ldots,n} \left(a_i + \langle s_i, p\rangle\right) - \max_{j=1,\ldots,m} \left(b_j + \langle t_j, p\rangle\right) > 0 \text{ for all } p \in D^+$$

$$\max_{i=1,\ldots,n} \left(a_i + \langle s_i, p\rangle\right) - \max_{j=1,\ldots,m} \left(b_j + \langle t_j, p\rangle\right) < 0 \text{ for all } p \in D^-$$

$\Longrightarrow$ union of polyhedral cones

# LINEAR AND RELU CLASSIFIERS

| | Linear | Piecewise linear / tropical rational / ReLU [B.-Loho-Montúfar |
|---|---|---|
| Parameters of same classification | Polyhedral cone | Union of polyhedral cones |
| Subdivision of parameter space | Hyperplane arrangement:<br>normal fan of a polytope<br>• Minkowski sum of 1-dimensional simplices (line segments)<br>• one summand per data point | Polyhedral fan:<br>normal fan of a polytope<br>• Minkowski sum of (n+m-1)-dimensional simplices<br>• one summand per data point |
| Local and global minima | All local minima of 0/1-loss are global minima | Local minima are not global minima |

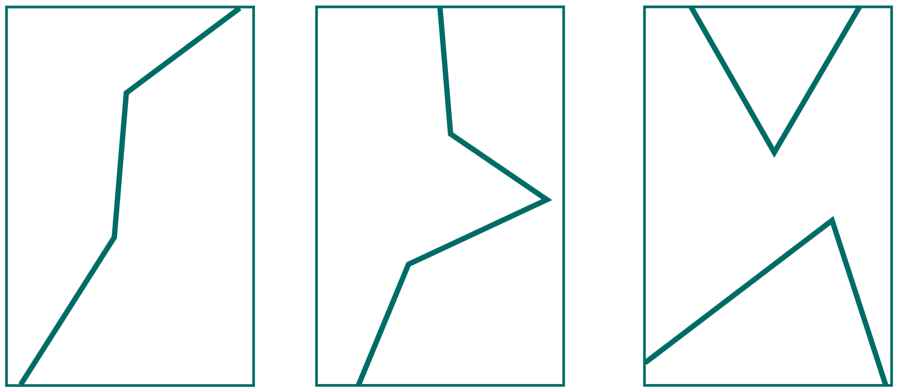
# LOCAL AND GLOBAL MINIMA

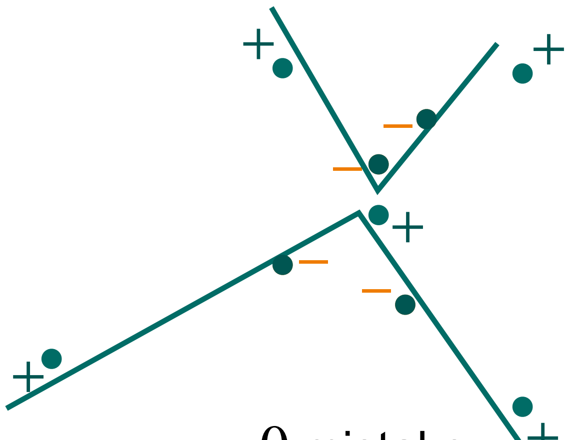
Classify 9 points in $\mathbb{R}^2$ in general position by piecewise linear functions (tropical rational functions) with $n = m = 2$ pieces.

Parameter space $\cong \mathbb{R}^{12}$, subdivided into $41680$ $12$-dimensional polyhedral cones.

Fix a labelling $D = D^+ \sqcup D^-$.

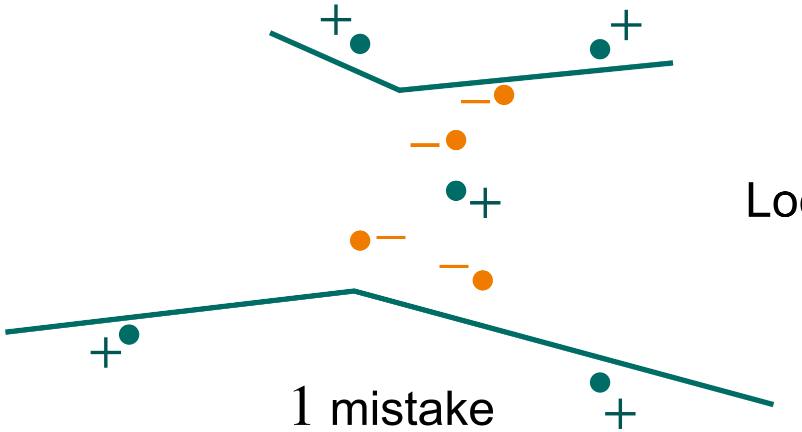
$16$ cones make $0$ mistakes, $8$ connected components
$304$ cones make $1$ mistake, $28$ connected components






Global minimum

0 mistake




Local minimum

$1$ mistake
but no path to a cone with $0$ mistakes
(through codimension $1$)

# LINEAR AND RELU CLASSIFIERS

| | Linear | Piecewise linear / tropical rational / ReLU [B.-Loho-Montúfar |
|---|---|---|
| Parameters of same classification | Polyhedral cone | Union of polyhedral cones |
| Subdivision of parameter space | Hyperplane arrangement:<br>normal fan of a polytope<br>• Minkowski sum of 1-dimensional simplices (line segments)<br>• one summand per data point | Polyhedral fan:<br>normal fan of a polytope<br>• Minkowski sum of (n+m-1)-dimensional simplices<br>• one summand per data point |
| Local and global minima | All local minima of 0/1-loss are global minima | Local minima are not global minima |

**Next Steps:**
Relate to fixed architectures of ReLU NNs

# PART II: ALGEBRAIC APPROACHES TO DEEP LEARNING

Guido Montúfar
UC LA | MPI MiS

Bernhard Reinke
MPI MiS

# TRAJECTORIES OF GRADIENT DESCENT

## Parametric model

The parametrization map takes parameter values $\theta \in \Theta \subseteq \mathbb{R}^p$ to functions $f(\,\cdot\,,\theta) \in \mathscr{F}$. We denote this map as

$$\begin{aligned} \mu :& \Theta \longrightarrow \mathscr{F} \\ & \theta \longmapsto f(\,\cdot\,,\theta)\,. \end{aligned}$$

The set $\mathscr{F}$ can be continuous functions from one set to another, for example.

## Loss function

Consider a loss function $\ell$ on $\mathscr{F}$. The corresponding loss on the parameter space $\Theta$ is defined as

$$\mathscr{L}(\theta) = \ell(\mu(\theta))\,.$$

## Trajectories of Gradient descent

Consider the trajectory of parameter values $\theta(t)$ for $t \geq 0$ of the dynamical system

$$\begin{aligned} \theta(0) &= \theta_0, \\ \frac{d}{dt}\theta(t) &= -\nabla \mathscr{L}(\theta(t)) \end{aligned}$$

# INVARIANCES OF TRAJECTORIES

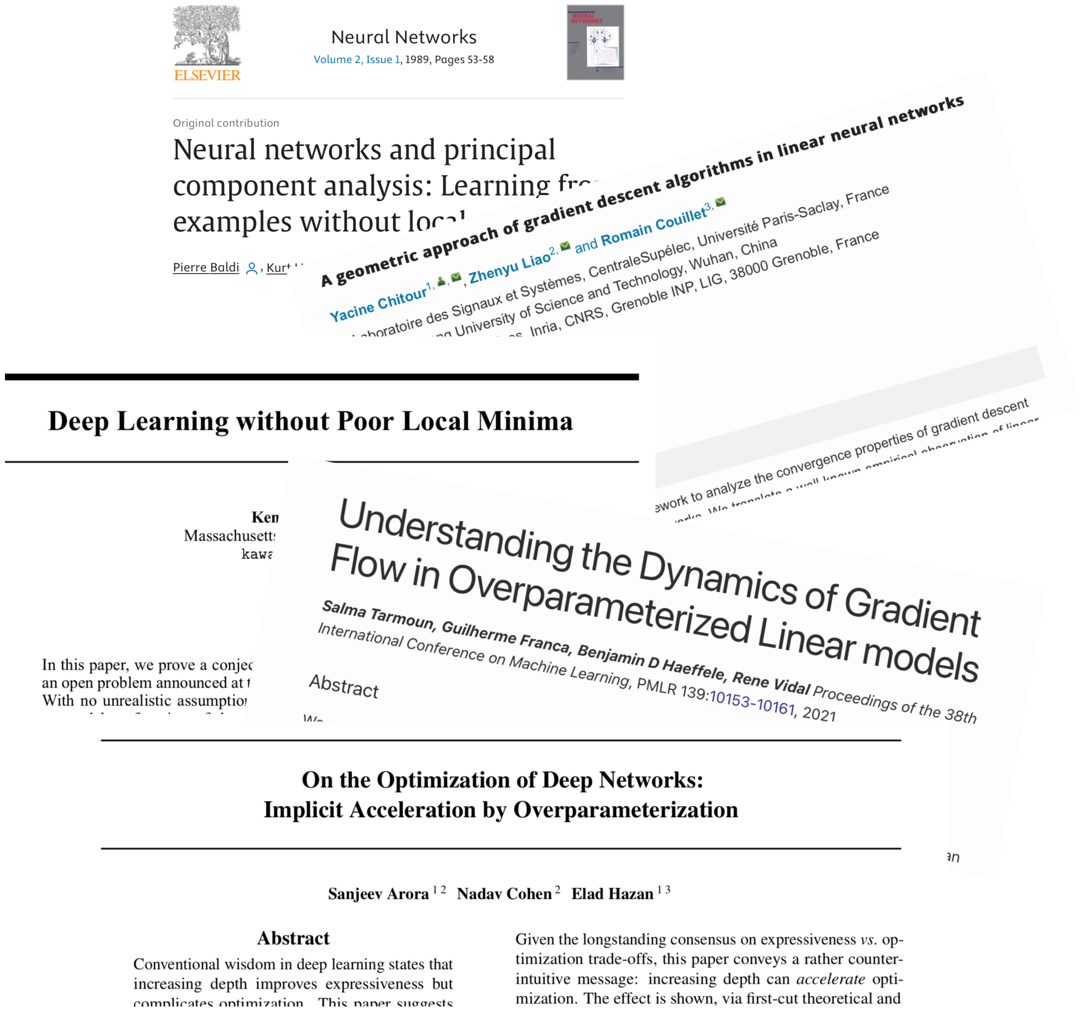
Consider the trajectory of parameter values $\theta(t)$ for $t \geq 0$ of the dynamical system

$$\theta(0) = \theta_0,$$
$$\frac{d}{dt}\theta(t) = -\nabla \mathscr{L}(\theta(t))$$

An invariance of the trajectory is a function

$$g : \Theta \longrightarrow \mathbb{R}$$

such that $g(\theta(t)) = 0$ for every $t \geq 0$.



# INVARIANCES OF TRAJECTORIES

## Short term goals

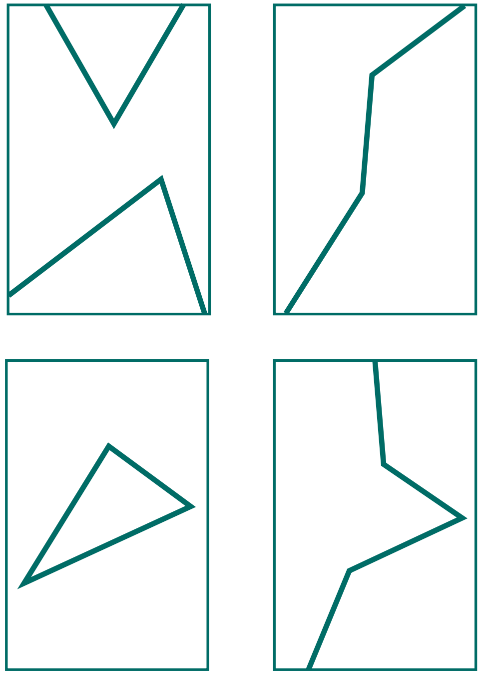
- For LNN: determine if the invariances previously known are complete (for quadratic loss)
- Design a systematic procedure to find such invariances

## Medium term goals

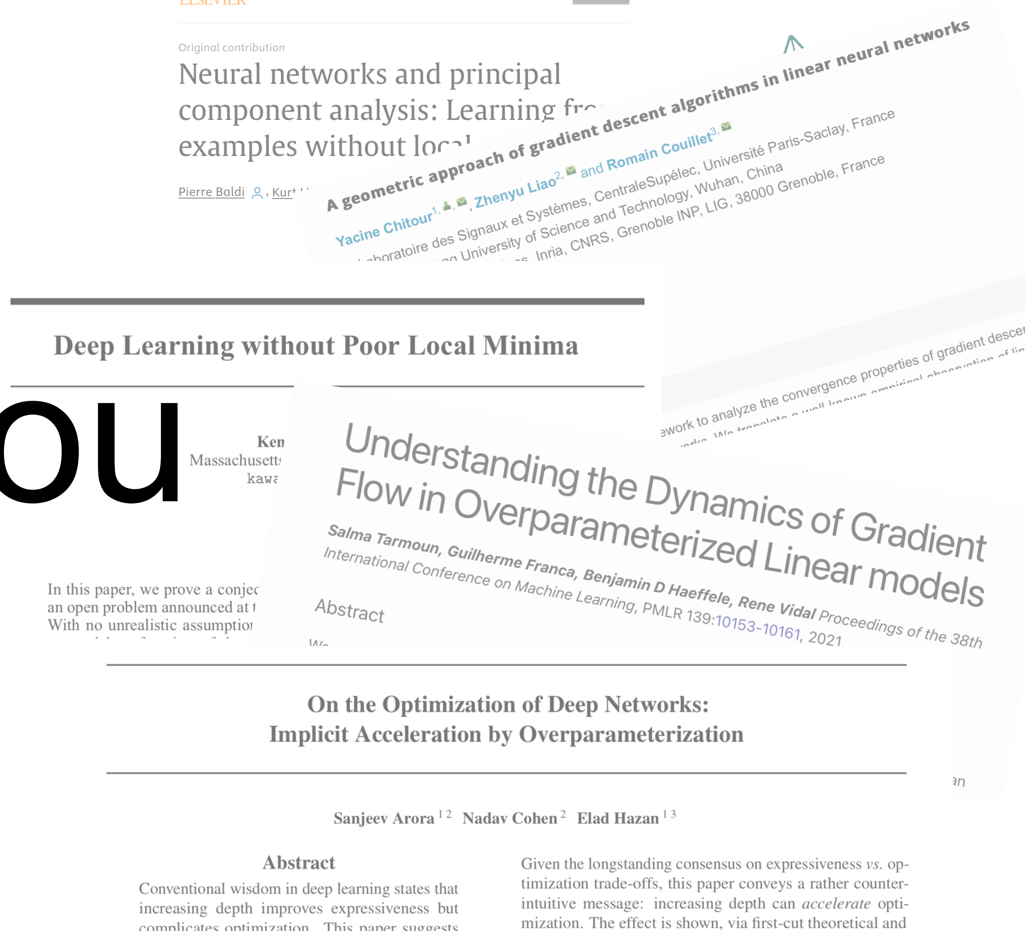
- Extend our methods to general loss functions
- Extend to optimization procedures with finite step size

## Long term goals

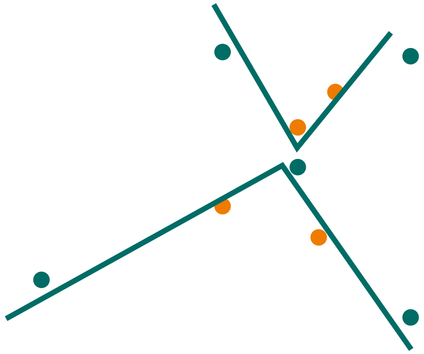
- Are extensions to sparsely connected linear networks or piecewise polynomially parametrized models possible?

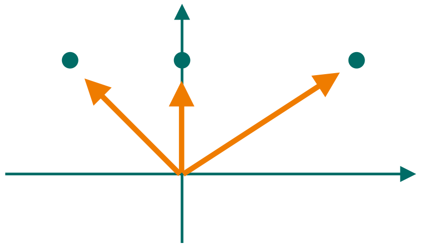


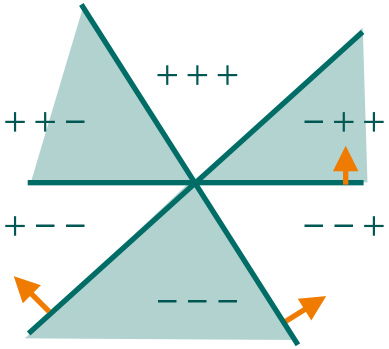
# Thank You



Original contribution
Neural networks and principal component analysis: Learning fr
examples without loc
A geometric approach of gradient descent algorithms in linear neural networks
Yacine Chitour, Zhenyu Liao and Romain Couillet
Deep Learning without Poor Local Minima
Understanding the Dynamics of Gradient Flow in Overparameterized Linear models
Salma Tarmoun, Guilherme Franca, Benjamin D Haeffele, Rene Vidal Proceedings of the 38th International Conference on Machine Learning, PMLR 139:10153-10161, 2021
Abstract
On the Optimization of Deep Networks:
Implicit Acceleration by Overparameterization
Sanjeev Arora Nadav Cohen Elad Hazan
Abstract
Conventional wisdom in deep learning states that increasing depth improves expressiveness but
Given the longstanding consensus on expressiveness vs. optimization trade-offs, this paper conveys a rather counter-intuitive message: increasing depth can accelerate optimization. The effect is shown, via first-cut theoretical and





+ + +
+ + −
− + +
+ − −
− − +
− − −